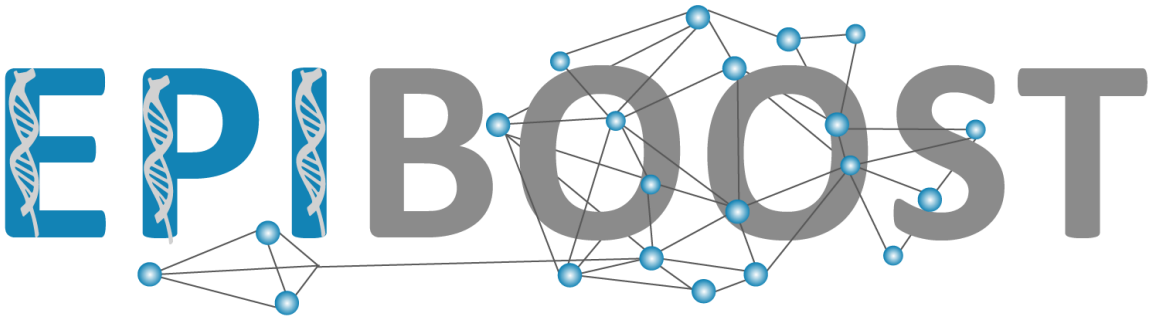
Deliverable D7
(D5.4)



EPIBOOST
BOOSting excellence in environmental EPIgenetics
(GA n. 101078991)

by
UAVR

universidade de aveiro theoria poiesis praxis

30 March 2023

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Union under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

Deliverable D7 (D5.4): Data Management Plan (DMP)

Authors: Inês Macário, Joana Pereira

Work package (WP)	WP5 Dissemination, Communication and Exploitation (DE&C)
Deliverable	D7 (D5.4)
Lead beneficiary	UAVR
Deliverable type	■ DMP
Dissemination level	■ PU (public)
Estimated delivery deadline	M6 (31/03/3023)
Actual delivery deadline	31/03/2023
Version	1.0
Reviewed by	Jana Asselman, Laia Navarro Martin

History of changes

Version	Publication date	Changes
1.0	31/03/2023	Initial version

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Union under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

Table of Contents

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

1. About this document

This document is composed of the Data Management Plan (DMP) for the project EPIBOOST and follows the "Horizon Europe DMP (version 1.0)" template provided by the European Union. As per its nature and the dynamics of its contents, the EPIBOOST DMP is a living document under continuous update as it will be recorded in the history of changes in the back cover page. An updated version of the DMP will be delivered formally in the second year of the project (Deliverable 13, due in March 2024) that will be elaborated in the light of any changes in data generation and outputs, providing extended detail through different sections covering data that have been managed by that time.

Following the template, this document is structured in seven main sections, with sections 3-9 corresponding to the DMP following the present and an introduction section. The DMP the EPIBOOST DMP starts by describing the data (section 3 – Data summary), including the purpose of data re-use and generation, data types and formats, size, origin and utility. Then the project's FAIR compliance is reasoned (section 4 – FAIR data). The strategy for handling and managing other research outputs follows (section 5 – other research outputs), considering the necessary adherence to FAIR principles when suitable. Finally, resources allocation (section 6), data security issues (section 7), Ethics, namely regarding personal data protection (section 8), are addressed, as well as any other issues and final remarks (section 9).

2. Introduction

The EPIBOOST project was funded under the call HORIZON-WIDERA-2021-ACCESS-03-01 for Coordination and Support Actions (CSA). The project Consortium is composed by the University of Aveiro (UAVR), the Coordinator, the University of Ghent (UGent, Belgium) and Consejo Superior de Investigaciones Científicas (CSIC, Spain), the advanced partners.

The present DMP constitutes a guideline complying to European and national legislation on the acquiring, handling, processing and archiving of data generated during the EPIBOOST project and in the project's afterlife, being updated in the second year of the timeline (Deliverable 13). The EPIBOOST DMP aims at allowing for a consortium-wide alignment concerning the project's data management processes, which comply with the EU Open Science policy framework, the EU General Data Protection Regulation (GDPR; Regulation (EU) 2016/679), and any national/institutional regulations or guidelines enforcing its provisions. The document is also naturally compliant with the rules for carrying out the Action as described in the Grant Agreement, concerning in particular Articles 13-17, including related specific rules detailed in Annex 5.

All partners will generate and handle diverse data within the scope of the project; thus, all are committed to the provisions detailed in the present DMP, which are either general or specific of each partner depending on internal data handling and protection routes, as well as resources available. In agreement, each partner generating or reusing research data will be responsible for their quality, organization and management, ensuring integrity and recoverability as necessary.

As detailed in the EPIBOOST Description of the Action (DoA; Annex 1 of the Grant Agreement), data management is addressed concerning implementation in WP5 – Dissemination, Communication and Exploitation –, within the scope of Task 5.5. The general principles ruling data management, originating the detail exposed in the present deliverable are also described in section 1.2 of the DoA – Coordination and/or support measures and methodology.

Following the "Horizon Europe DMP (version 1.0)" template provided by European Commission, the EPIBOOST DMP is structured in seven sections after the present introduction.

• Data summary (section 3), where the data are summarised. This includes the identification of the data that will be re-used and that will be generated by the project, as well as its characterisation

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

regarding the purpose of re-use or generation in relation to the project objectives, size, type and format, origin and utility.

• Compliance with the FAIR (Findable, Accessible, Interoperable and Re-usable data) and open data sharing by default ambition of the EU's open science policy for the results of EU-funded scientific research (section 4). As directly interpretable, this section presents the EPIBOOST strategies to render the project data Findable, Accessible, Interoperable and Re-usable during the project and in its afterlife.

• Management of other research outputs (section 5), considering the necessary adherence to FAIR principles when suitable. This will be fundamentally focussed on the workflows and protocols that will be developed for excellent research in environmental epigenetics, as well as training outlines.

• Allocation of resources (section 6) as needed to render EPIBOOST data and other research outputs FAIR, including the reasoning of any costs, staff responsibilities and long-term preservation.

• Data security (section 7), where the procedures employed by each partner to ensure that all data generated during the project will be securely stored/archived will be described, including the necessary appraisal of data recovery suitability, transfer of sensitive data provisions, and long-term preservation and curation when applicable.

• Ethics (section 8), where personal data protection strategies will be addressed for covering the data that will be processed entailing sensitive information about participants in EPIBOOST activities.

This plan addresses the main issues concerning EPIBOOST open science approach. It conveys the commitment of the project towards the sharing of research outputs and other project outcomes as open as possible and as early and widely as possible, while sufficiently closed to respect intellectual property rights and the privacy rights of participants. FAIR principles will guide the sharing and dissemination efforts, making use of existent public and institutional routes as recommended in the Horizon Europe Open Science framework.

3. Data summary

The overarching aim of EPIBOOST is to stimulate EU research excellence through capacity building of the widening UAVR coordinator and concomitantly to strengthen the Consortium towards a world-class level for efficiently tackling the challenges of the proposed incorporation of epigenomics into regulatory Environmental Risk Assessment (ERA) frameworks. Meeting the call conditions and the objectives of the project, both scientific and capacitation activities will generate and/or require the use/re-use of data, as described in the sub-sections below and summarized in Appendix A.

3.1 Data re-use

As a capacitation Action, EPIBOOST is mostly focused in learning by doing; moreover, the research foreseen in the project (WP1) comprises a strong experimental component that necessarily generates new data instead of reuse of existing data. Still, in the present first version of the DMP, the reuse of genomic data (section 3.1.1) and personal data (sections 3.1.2) is foreseen.

In WP1 (Preparatory research project: epigenetic responses of model organisms to relevant environmental contaminants), the reuse of data on the sensitivity of test organisms to the defined model contaminants was considered through Task 1.1 (Experimental planning and resources organization). This would accelerate the implementation of the main experiments in Task 1.2 (Exposure of model organisms to selected contaminants), overcoming preliminary experiments to define exposure levels. However, laboratory conditions and specific strains used can constrain the

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

sensitivity of aquatic organisms. Therefore, the reuse of existent data for the purpose of establishing treatments in the project experiments was found inadequate and preliminary experiments to generate own and specific data were preferred. The exception to this workflow is the seabass, a vertebrate that is handled under a strict ethics framework. In this case, exposure levels will be based on indications provided by the literature, both considering toxicity assessment and environmental concentrations. As this collected information will not be used systematically to validate any hypothesis, neither will be found as assembled evidence in the form of single datasets, EPIBOOST does not consider it as formal data to be reused and treated as such in the present data management plan.

3.1.1 Genomic data

Available genomic data and metadata on the species that will be used for research and hands-on research training (two microalgae, two microcrustaceans and two fish species) will be reused primarily in tasks related to bioinformatics and data analysis, within WP1 (Task 1.5: Bioinformatics and data analysis) and WP2 - Hands-on capacitation of in-house researchers for nurturing research on epigenetic responses to environmental stressors in aquatic ecosystems (Task 2.4: Exchange and sessions on bioinformatics & data analysis). Additionally, these data will also be reused to support the design of primer sequences needed in targeted methylation and gene expression approaches as foreseen in WP1, Task 1.3. The availability of good quality genomic data as required in EPIBOOST concerning all species addressed was already confirmed from NCBI (National Center for Biotechnology Information) GEO (Gene Expression Omnibus) and SRA (Sequence Read Archive); still, other sources such as ENA (European Nucleotide Archive) will be additionally considered as the information is needed and used if better quality data is found therein.

Raw genomic datasets and corresponding metadata are typically available as textual data in non-proprietary, standard formats used for sequence reads such as FASTQ and TXT/GFF, BAM, compressing SAM files, and HDF files (.fast5). The genomic data for each species that will be reused by EPIBOOST range in size up to 10 GB, depending on the species, the level of processing of the information required and/or on the extension of each dataset.

3.1.2 Contacts of the Coordinators of sister projects

Within the context of WP5 – Dissemination, Communication and Exploitation (Task 5.2 – Stakeholder mapping and engagement), the contact with the Coordinators of other Twinning projects funded under the same call as EPIBOOST (sister projects) was foreseen. Besides being a natural target of the dissemination and communication strategy, members of this community are an expectedly active participants in a Focus Group aiming at the co-creation of a good practices guide on the management of Twining projects. This outcome contributes to the tackling of the EPIBOOST capacitation aims concerning science management and administration, under Specific Objective 4: *To strengthen Science management and administration skills of UAVR, concomitantly improving the capacity of the current Consortium to grow and embrace wider funding opportunities*.

Following the Coordinator's KO day – TWINNING (22/09/2022) and the requests by different projects, the Research Executive Agency shared among all Coordinators a dataset with the identification and contacts associated to all projects funded under the call. This sharing was made through the CIRCABC (Communication and Information Resource Centre for Administrations, Businesses and Citizens) platform, though a downloadable tabular file (.xlsx; 41 KB) containing textual information: Proposal Number, Proposal Acronym, Panel, Title of the Project, Primary contact first name, Primary contact family name, Email, Country. These data are processed for the creation of mailing-lists filtered for dedicated contacts (surveying, meeting scheduling) in the proprietary MS®Excel software (licenced to UAVR, the data handling partner in this case), then converted in .csv files.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

3.2 Data generated by the EPIBOOST project

EPIBOOST will generate research data as part of the work developed under WP1, as well as will collect and handle data that in some cases include personal data, concerning EPIBOOST capacitation activities and its participants. The sub-sections below detail on types and formats (3.2.1), the purpose of generating/collecting these data (3.2.2), size (3.2.3), origin (3.2.4) and utility (3.2.5) of the data generated/collected. An overall summary of the information given is provided in Appendix A.

3.2.1 Data types and formats

- **Data types**

Data generated in EPIBOOST will generally be of the following types, depending on the concerned activities. As to the content, we will produce numerical, textual or graphical; as to the nature, data will be raw or processed, qualitative or qualitative.

Research data that entail genomic information (nucleic acids quantity and quality, DNA methylation, gene expression) are numerical and/or textual, always quantitative; raw and processed data will be managed through the project course (e.g., raw datasets will be merged and/or analysed and derive processed datasets), except for nucleic acids quantification, which constitute support data (see section 3.2.2), and will be managed as raw data only. Research data compiling on phenotypic outcomes of experiments are additionally graphical, this being applicable in some cases to raw data (e.g., movement tracking in behavioural assays), but especially to processed data, complying to widely used descriptive statistics standards.

Data relating to attendants and evaluation of EPIBOOST courses, summer schools and workshops open to external participation, are textual, numerical and graphical, depending on the surveyed aspect and the reporting needs. Both qualitative and quantitative data will be managed, depending on reporting needs/format and associated dissemination and communication strategies/flows. Owing to the origin (section 3.2.4) and purpose (section 3.2.2) of the data, raw and processed data will be handled and managed, with graphical data being available only after processing. The same data types apply to the information collected for the activity of the EPIBOOST Focus Group on Twinning management, following the reuse of data concerning sister projects (section 3.1.2).

- **Data formats**

Either research data or data generated from EPIBOOST capacitation activities will be (collaboratively) handled by the team and shared within the Consortium, then shared externally when applicable, as detailed further in section 4 of the present document.

Accounting to the partner's institutional software resources, proprietary data formats will be often used for data management during internal handling, e.g., MS® Excel (.xls/.xlsx) and MS® Word (.doc/.docx). Despite the proprietary nature of these formats, all partners are granted access to compliant software either locally or online; also, these are widely-used formats with high degree of compatibility with open-source equivalents, which are often accepted in many FAIR compliant data repositories. Other formats applying to data collection and processing are non-proprietary, namely:
a) Text formats that are standard yet specific of sequence reads (.fq, .fast5,.dta), both covering for raw and processed data (e.g., functional annotation and/or counting), that will be collected from Task 1.3 and processed in Task 1.5 for analysis.
b) Text formats (.txt, .csv) to be used in research data for analysis (bioinformatics and statistics), as well as concerning personal data in mailing-list assembly for example.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

c) Light image formats (.jpg, .tiff), allowing an easy internal sharing and incorporation in working text files, both concerning research data and data linked to capacitation activities.

Non-proprietary data formats will be exclusively used for data sharing and long-term preservation when applicable. Concerning research data, any MS® Excel or MS® Word final (fully processed) datasets will be converted into comma-separated values (.csv), plain text (.txt) or PDF (.pdf) formats; raw and processed sequencing data will be converted into if necessary and preserved in non-proprietary formats as required by the data repositories used in EPIBOOST (.fq, .fast5, .dta, .txt, .csv). In all cases where data sharing is made via reports (deliverables, scientific publications, dissemination items), the Adobe Portable Document Format (PDF/A, PDF) (.pdf) will be used.

3.2.2 Data generation purposes

The generation of research data in EPIBOOST is a direct consequence of its ambition to contribute to the knowledge pool in the field of environmental epigenetics, towards the clarification of Adverse Outcome Pathways and the development of epigenetic biomarkers, both assisting the improvement of Environmental Risk Assessment frameworks. In this context, research data generation will occur mostly through the implementation of WP1 (research project) and WP2 (hands-on training concerning the research developed in WP1). These two work-packages tackle the first specific objective of the project: *To develop research protocols for the comprehensive assessment of epigenetic effects and its consequences in aquatic organisms, leveraging UAVR scientific skills through hands-on training and the research profile of the Consortium in the field of Environmental epigenetics through the production of excellent knowledge*.

In this context, and as detailed in Deliverable 5 (Detailed Experimental Plan), a series of experiments will be set up, originating records on responses of 6 aquatic organisms (two microalgae, two microcrustaceans and two fishes) to exposures to three model contaminants. The expected records will originate diverse data with specific purposes as follows:

• **Nucleic Acids quantity and quality records concerning different species, before and after exposure experiments**. Before experiments, trial extraction and purification of nucleic acids (DNA and RNA) with the consequent assessment of yield quantity and quality (Task 1.2) is critical to appropriately frame the experimental design concerning the number of test organisms required for a successful downstream analysis of DNA methylation and gene expression following definitive experiments. After exposure experiments these records are necessary, as part of Task 1.3, for confirmation on the suitability and feasibility of downstream applications to analyse DNA methylation and gene expression.

• **DNA methylation datasets concerning different species and challenges experimentally imposed to the organisms**. Considering the focus of the project on environmental epigenetics, DNA methylation in particular, the purpose of generating these data is obvious. These data are the basis for the understanding of the molecular initiating events driving negative outcomes of exposure to the environmental contaminants selected in EPIBOOST. Gathered in Task 1.3, the data will be handled and processed in Task 1.5 (supported by Task 2.4) for analysis.

• **Transcriptomic datasets concerning different species and challenges experimentally imposed to the organisms**. As the epigenome regulates the transcriptome, the generation of these data is critical to elucidate the link between stressor-induced DNA methylation and effects in gene expression for the different species. This link provides the confirmation that DNA methylation can initiate an adverse outcome pathway by modulating the transcriptome, hence onsetting the routes for a phenotypic outcome. Gathered in Task 1.3, the data will be handled and processed in Task 1.5 (supported by Task 2.4) for analysis.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

• **Phenotypic responses of challenged organisms, concerning several species and 2 or 3 environmental contaminants**. Several phenotypic endpoints of different type (sub-cellular, growth, behaviour) have to be monitored to clarify whether initial DNA methylation changes actually translate into phenotypic changes, since these later are the final outcome of exposure that can elucidate on the hazardous potential of environmental contaminants. Confirmed phenotypic effects are evidence of an adverse outcome pathway, and when consistent with a given DNA methylation pattern, expose potential DNA methylation biomarkers. These data will be generated and processed in Task 1.4.

Other data that will be generated during EPIBOOST concern capacitation activities and are therefore related to specific objectives of the project in this regard, namely: *to contribute on the building of a sustained European critical mass to leverage environmental epigenetic research to the best world-class standards through broader attracting and training of young researchers* (specific objective 2); *to strengthen research resources management of UAVR, appropriately tuning key infrastructures to support world-class projects in the field of environmental epigenetics* (specific objective 3); *to strengthen Science management and administration skills of UAVR, concomitantly improving the capacity of the current Consortium to grow and embrace wider funding opportunities* (specific objective 4). In this context, different data will be generated with specific purposes as follows:

• **Data concerning participants and evaluation of Advanced Courses**. These Courses are part of WP3 (Task 3.1 - Advanced courses on environmental epigenetics), targeting early-career researchers and in particular PhD students. Their training at these early stages is a paramount step towards raising interest for the field, feeding the continuation and growth of European research in environmental epigenetics. Data in this case consist of the collection of opinions and views of participants regarding course quality and utility, which will allow both an assessment of the EPIBOOST performance and its improvement from the first edition (2023) until the last edition (2024). In addition, personal data will be collected and handled concerning registration in the courses and authorization for the capture of image, sound and video, for reporting (Deliverable 16 - Report on the implementation of advanced courses) and dissemination purposes according to the project Dissemination and Communication strategy (Deliverable 6 - Dissemination and Exploitation plan, including Communication).

• **Data concerning participants and evaluation of Summer Schools**. These Summer Schools are part of WP3 (Task 3.2 - Summer Schools on environmental epigenetics), targeting early-career researchers. These Summer Schools have a similar training component as Advanced Courses, but complemented with networking opportunities that are naturally stimulants of future investment in collaborative research in the field of environmental epigenetics anchored in societal challenges. Data in this case consist of the collection of opinions and views of participants regarding Summer Schools' quality and utility, which will allow an assessment of the EPIBOOST success in this Task. In addition, personal data will be collected and handled concerning registration in the Summer Schools and authorization for the capture of image, sound and video, for reporting (Deliverable 21 - Report on Summer Schools implementation) and dissemination purposes according to the project Dissemination and Communication strategy (Deliverable 6 - Dissemination and Exploitation plan, including Communication).

• **Data concerning participants and evaluation of Short Courses**. These Short Courses (three editions – one per year) are part of WP3 (Task 3.3 - Short-courses in scientific meetings), targeting early-career researchers attending scientific conferences in the field of environmental risk assessment. They were specifically planned to provide a short overview on the importance of epigenetics in modern environmental risk assessment techniques and provide a quick snap-shot of the research possibilities in the field. Data consists of the collection of opinions and views of participants regarding Courses' quality and utility, allowing an assessment of the EPIBOOST success in Task 3.3 and the consequent

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

adjustments in further editions. Depending on the interaction with conference organizing committees and on their internal procedures, personal data may be collected and handled concerning registration and authorization for the capture of image, for both reporting (Deliverable 20 - Report on the implementation of SETAC courses) and dissemination purposes according to the project Dissemination and Communication strategy (Deliverable 6 - Dissemination and Exploitation plan, including Communication). Still, the availability of these data will be constrained by the Conference policy on the sharing with course organizers.

• **Data concerning participants and evaluation of the workshop on ethically compliant use of animals in research**. This workshop is part of WP4 (Task 4.1 - Training on ethically compliant use of animals in research) and will be open to participation of researchers and technicians external to the EPIBOOST team. It meets the project objectives related to the capacitation of staff and infrastructures at UAVR to develop excellent research using animals, under the best ethics standards concerning animal housing and experimentation. Data to be collected and handled consist of the collection of opinions and views of participants regarding training quality and utility, which will allow an assessment of the EPIBOOST success in this Task. In addition, personal data will be collected and handled concerning registration and authorization for the capture of image, sound and video, for reporting and dissemination purposes according to the project Dissemination and Communication strategy (Deliverable 6 - Dissemination and Exploitation plan, including Communication).

• **Personal data, views and opinions of participants in the Focus Group on the management of Twinning actions**. As described in section 3.1.2 above, a Focus Group on the management of Twinning actions will be created to meet EPIBOOST specific objective 4, within the context of WP5 (Task 5.2). The surveying of the community of sister projects will be regularly made via digital forms, meetings and collaborative work, originating data that will be processed to support the co-creation of a good practices guide on the management of Twining projects. In addition, personal data will be collected and handled concerning adherence to the Focus group with differential levels of involvement, as well as authorization for the capture of image, sound and video during meetings, for reporting and dissemination purposes according to the project Dissemination and Communication strategy (Deliverable 6 - Dissemination and Exploitation plan, including Communication).

3.2.3 Data size

The expected size of data will be continuously estimated during the whole period of the project (3 years) due to the variability of data types and quantity. At this early point in the timeline, the data size reasoned is hence merely indicative, and the Coordinator will ensure availability of resources for local storage of the data even if the currently reasoned sizes are severely underestimated, considering available active storage resources for the project (5 TB; see section 4.2). Appendix A summarises the expected size for the data (work package and task specific), accounting to raw, processed and final data, merging several datasets within each data category as shortly described. The size foreseen for data concerning participants and evaluation of different activities considers the storage of captured image, sound and/or video records, which should be considered personal, sensitive data (see section 8), for use in reporting stages and dissemination efforts.

As a general guideline, we are expecting to generate around 2 TB of data within EPIBOOST, composed by datasets that range the KB figure (e.g., contact lists concerning specific stakeholders such as the coordinators of sister projects) to datasets that may size up to 1.5 TB as it is the case of raw and processed data concerning DNA methylation and transcriptomic assessment accommodating 6 species, 2-3 stressors as exposure treatments and different assessment techniques (including but not necessarily whole genome sequencing).

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

3.2.4 Data origin/provenance

Research data generated by EPIBOOST will originate from experiments carried out during WP1 with the selected species – marine and freshwater equivalents from standard groups in aquatic ecotoxicology, i.e., microalgae, microscrustaceans and fish. These data are hence of an experimental nature. Given the intense collaborating foreseen in the project due to its capacitation objectives, research data will be generated by the three partners, UAVR, CSIC and UGent. As per the structure of the research project and the implementation options, the largest part of research data will be generated at UAVR. CSIC will be generating specific datasets concerning fish and UGent will be specifically active concerning microalgae and microcrustaceans.

Data concerning training and capacitation activities will originate from surveys, applied by each of the three partners depending on the implementation strategy. Accordingly, data provenance is assigned to: UAVR for Advanced Courses that will be always implemented therein (WP3; Task 3.1); CSIC and UGent for Summer Schools, one held at UGent in 2023 and the other held at CSIC in 2025 (WP3; Task 3.2); UGent for Short Courses in Scientific Conferences, considering that UGent is the leader for this Task (WP3; Task 3.3), pending agreement in sharing by Conference organizers; UAVR for the workshop on ethically compliant use of animals in research that will be organised therein likely in a hybrid format (WP4; Task 4.1); UAVR in what concerns the Focus Group on the management of Twinning actions (WP5; Task 5.5), here with data originating from meetings adding to data originating from surveys.

3.2.5 Utility of Data outside the project

Project activities and research results are planned to reach the previously identified stakeholders which will be beneficiaries of the project. Accordingly, to the stakeholder's map developed (Deliverable D4), the project identified six broad categories of stakeholders: 1) early-career researchers, 2) the EU twinning community, 3) the scientific community, 4) policy makers and regulators, 5) the private sector, and 6) the society in general, including the educational community and the general public.

Research data that will be generated during the course of WP1, supported by WP2, will be useful for the scientific community, including early career researchers, who can benefit from the shared raw datasets, their processed versions and published papers incorporating their interpretation. The availability of these data will support the onset of new, more finely tuned research in the field of environmental epigenetics, promoting the development of this research field. Moreover, the data in its processed and interpreted versions will be useful to regulators and policy makers dedicated to environmental assessment and protection. This relates to the central aim of the EPIBOOST project concerning research excellence that reflects the willing to contribute for the incorporation of epigenetic biomarkers in risk assessment frameworks and broadly the consideration of epigenetic mechanisms as molecular initiating events in Adverse Outcome Pathways. In such arenas, regulators and policy makers are key actors needing data to support their actions/activities.

Science managers and the EU Twinning community at the EU level are the target audiences of the data reflecting views and opinions of participants in the Focus Group on the management of Twinning projects. These data will be synthetised in one co-created good practices guide on the management of Widening-Twinning projects. This guide will expectedly be a well-tuned, experience-based tool that can be exploited in the long term by Coordinators of future Twinning projects.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

4. FAIR data

The management of EPIBOOST data (see Appendix A for an overview) will follow the FAIR protocol, the Consortium being committed to the best extent to mandatory and recommended open science practices, advocating openness by default as early and widely as possible while respecting intellectual property rights (IPR). Sharing of the research data not protected by IPR that will be made open in the long-term will occur at the latest by the end of the project if not before, upon publication of the respective scientific manuscripts. Each partner generating data will be responsible for their quality, organization and management, ensuring integrity and recoverability as necessary, as well as complying with the FAIR protocol. Such compliance covers the methods for rendering the data open when relevant, but also the handling and sharing of the data within the Consortium during the project implementation.

4.1 Making data Findable, including provision for metadata

Findability will be ensured by using the best standards in data identification, including the association of automatically generated globally unique and persistent identifiers, as well as in associating the dataset to standard descriptive metadata and corresponding keywords.

4.1.1 Data identification

Research data generated by the project will be deposited in repositories (see section 4.2) that automatically associate a standard persistent identifier (PID) upon or soon after upload for archiving. In EPIBOOST, three PIDs and one stable identifier will be used depending on the data itself, on the document carrying the data and/or on the repository where the data will be archived:

• DOI. This PID will be used for data deposited in Zenodo or in institutional repositories using this PID. This is the case of data regarding phenotypic responses recorded in WP1. In addition, as processed data will be integrated in scientific publications, they will also be identified by DOI in this context.

• Handle. This PID will be used for the identification of processed data included in articles and scientific documents other than articles, archived in institutional repositories using this PID. It applies for example to PhD and MSc thesis, conference communications, policy briefs developed under the scope of the project deposited in the UAVR repository. The Handle System allows assigning, managing, and resolving "handles" for digital objects and other resources on the Internet.

• ORCID. This PID will be used to identify the authors of the data whenever repositories allow the association of this PID to the (meta)data being archived.

• Accession number. This identifier, as well as their linked identifiers for versioning and subsets will be automatically associated to DNA methylation and transcriptomic datasets generated in WP1, since these will be archived when ready in the repository Gene Expression Omnibus (GEO; see section 4.2.1). GEO accession numbers cover the complete data/database submitted, are unique and are stable, even if updates/revisions are made to the record. While this is not a standard PID system, as per the feasibility and establishment of the repository in the field, it is recommended by many scientific publications, often recognised as a permanent identifier.

Concerning data derived from capacitation activities, the results of the integration of views and opinions of participants in the Focus Group on the management of Twinning actions will be enclosed in a document (Good Practices Guide on the management of Twinning actions) that will be identified

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

by DOI. Other data related to capacitation activities are internal and not entitled to long-term preservation requiring a PID.

During active storage stages before open external sharing and for internal data not entitled to long-term preservation (see section 4.2 for accessibility details), findability by the EPIBOOST team will be promoted by benefiting from a structured organization of files within shared folders that correspond to the different work packages (WP) and tasks, as well as by adhering to a naming convention for file names. As a general guideline, file names in the EPIBOOST shared documentation will be given so that they can be machine readable (especially in the case of genomic datasets that will be heavily processed within the project), human readable and allowing default ordering. Depending on the specific requirements that may be found during the project in this regard, the EPIBOOST convention may be further elaborated, but at present, file names are established by an ordering number (e.g., date following the ISO standard YYYYMMDD), followed by the subject (short description of the content), the workpackage identification, authors initials, version (two digits) and extension, with eliminated spaces throughout the whole name of the file.

4.1.2 Metadata

For all data entitled for open long-term preservation, findability will be supported by rich associated metadata. All repositories that will be used (see section 4.2.1) comply to disciplinary and general metadata standards that are widely recognised by the research community and/or the public. Metadata will be elaborated in English and will be structured describing, explaining, locating and/or facilitating data retrieval, use or management.

Generically, the DataCite metadata schema will be followed when depositing data in the defined repositories. Mandatory DataCite properties, namely Identifier, Creator, Title, Publisher, Publication, Publication year and Resource type (controlled list, the most likely options being BookChapter, Book, ConferencePaper, ConferenceProceeding, DataPaper, DataSet, Dissertation, JournalArticle, Report), will always be provided. Recommended DataCite properties will also be incorporated in EPIBOOST metadata such as the Subject, Contributor(s) identified by the ORCID PID whenever possible, Date, Related Identifier for related data or versions and Description. Among optional DataCite properties, Version, Rights and Funding Reference (fulfilling the Grant Agreement requests) will be included. These standards will be followed for EPIBOOST data and outputs other than sequencing data (see below), deposited in public and institutional repositories (see section 4.2.1) that gather metadata on the base of embedded forms upon data submission.

Domain specific metadata standards will be followed additionally for sequencing data that will be produced in EPIBOOST. In this case, the MINSEQE (Minimum Information about a high-throughput nucleotide SEQuencing Experiment) standard (DOI 10.5281/zenodo.5706412) will be followed, considering that adherence to the MINSEQE guidelines will improve integration of multiple experiments across different modalities, thereby maximising the value of high-throughput research. Metadata within this standard framework includes critical information such as: the detailed description of the species, type of samples (tissue or whole body), and the experimental variables (contaminants and their concentrations/doses as tested); information about the sequence read data and base-level quality scores for each assay, the FASTQ format being recommended, with a description of the scale used for quality scores; the summary of the processed data, enlightening the conclusions in related publications based in the data, and descriptions of the data format; general information about the experiment and sample-data relationships, including experimental goals, contact information and associated publication(s); essential experimental and data processing protocols, including methods for nucleic acids isolation and purification prior to sequencing, instruments used, library preparation, labelling, amplification, alignment and quality control, plus analysis pipelines. As per the request of

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

the applicable repository (see section 4.2.1), metadata for sequencing data will be provided by fulfilling and submitting a template (generally a .xls file) for verification and curation by the repository before approval of the data submission.

4.1.3 Metadata keywords

Metadata associated to EPIBOOST data openly shared will be provisioned with specific search keywords rationally defined to improve findability/discoverability as much as possible. In fact, this is a requirement of the repositories elected by the project for deposit and long-term preservation of research data and other research outcomes. Ontology standards will be followed to guide the definition of keywords to improve interoperability as described in section 4.3, but as a general guideline, we will structure keywords definition within major ontology categories that will be always covered to properly position each dataset and the project focus within the biosciences/environmental sciences landscape: epigenetics, gene expression, phenotype (detailed to the entities involved); organism group and species, so that the study can be immediately identified as a non-human study; environmental compartment, specifically distinguishing freshwater and marine focus; exposure, with specific reference to each environmental contaminant involved, treatments and endpoints.

4.2 Making data Accessible

EPIBOOST supports open access to scientific information, aiming to comply with the best standards of the Horizon Europe Open Science framework. All research data generated by the project will be made openly available for open access at least at the time of publication in the permanent literature by achieving in trusted repositories that ensure accessibility in the long-term (see below). Long term preservation of data is essential to ensure future scientific complements to EPIBOOST, concerning to improve knowledge on molecular biomarkers as tools for ecological risk assessment frameworks. The creation of scientific databases allows the data to be reanalysed in the future in the light of technical and conceptual advances. Also, long term preservation of data allows to avoid duplication of studies, saving time, effort, and more important model animals.

However, the management of data is not restricted to its final archiving and hence accessibility needs to be appraised also for initial storage, organization and processing of the project data. At these stages, accessibility is naturally framed on a within-Consortium basis, ensured through a secure Onedrive platform hold and managed by the Coordinator (5 TB storage capacity). All researchers as confirmed by work package leaders will be given access to research data being generated and processed within this platform. For data deriving from surveys where sensitive information and personal data can be collected, access within the consortium will be restricted as described in section 8. The ultimate responsibility for access is hold by the project scientific coordinator at UAVR, Dr. Joana Pereira.

4.2.1 Repositories

Accessibility of research data will be granted as per the use of repositories providing http(s) or ftp communication protocols for globally implementable (meta)data retrieval. Affiliation and contact information will always integrate descriptive metadata so that datasets can be recovered in the long-term if for any unexpected reason their public availability becomes no longer sustained. All the non-institutional repositories used by EPIBOOST are catalogued in the registry of Research Data Repositories (Re3data - https://www.re3data.org/), support free, open access to the repository and

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

open access to the data. All repositories ensure long-term preservation of (meta)data and/or publications, as well as render them openly accessible under the conditions defined by submitters (e.g., embargoes until publication), under defined identifiers resolving digital objects (see section 4.1.1).

The following non-institutional repositories will be used to deposit research data generated by EPIBOOST:

● **Gene Expression Omnibus** (GEO; re3data.org record under DOI 10.17616/R33P44)
GEO is an FTP-based, curated, reliable and trustworthy public functional genomics data repository, governed by the National Center for Biotechnology Information (NCBI) and the National Institutes of Health, U.S. National Library of Medicine (NIH), supporting and encouraging MINSEQE (Minimum Information About a Next-generation Sequencing Experiment) compliant data submissions, accepting array- and sequence-based data. As such, and considering that it is one of the most used disciplinary repositories for sequencing data, GEO will be used to deposit DNA methylation and gene expression datasets generated in the project (see Appendix A). Although it is not yet indexed by OpenAIRE, GEO is covered by Clarivate Data Citation Index within Web of Science, which ensures many of the supportive features of OpenAIRE regarding the continuous reporting of (meta)data and publications. Appropriate alternatives that are OpenAIRE compliant are still unavailable. This is the case of another well quoted (although of less widespread use and with shorter ranges of data types accepted compared to GEO) repository for nucleotide sequences, ENA – European Nucleotide Archive (10.17616/R3HW3J). GEO allows versioning and their system is integrated in a way that records cross reference each other, so that the users can retrieve the current or earlier versions of each sequence record.

● **Zenodo** (re3data.org record under DOI 10.17616/R3QP53)
Zenodo is the OpenAIRE repository hosted by CERN. It is a general-purpose repository, where EPIBOOST will deposit data concerning the phenotypic responses of the challenged organisms (several marine and freshwater species and different environmental contaminants as defined in the project work plan – WP1; Appendix A). Zenodo is fully compliant with the best practice guidelines for Horizon Europe Open Science, being a trustworthy institutional repository that supports long-term data preservation safely and feasibly under DOI assignment, which also ensures unique citation stability. Versioning is fully supported by DOI (one per version linked in machine-readable metadata) and Concept DOI (semantically linking all the per-version DOIs) assignment, as well as immediate accessibility or lift-embargo depending on the submitter requests.

Besides their deposit per se in the above repositories, processed data will be also made available through scientific publications, policy briefs, academic theses and conference communications. At least upon publication, all scientific articles containing EPIBOOST research data will be made immediately accessible through adherence to gold open access principles. In addition, they will immediately be deposited in institutional repositories following the best practice recommended/demanded at the partner institutions. These repositories are as follows.

● **RIA - Institutional repository of UAVR** (https://ria.ua.pt/?locale=en)
RIA is an information system that captures and preserves the research outputs of UA scientific community and make them available over the Web in the long-term, increasing visibility and impact. It ensures open access to the full text version of documents upon authors' authorization as it will be the case in EPIBOOST publications. RIA integrates the Portuguese Open Access Science Repository (RCAAP) and is harvested by OpenAIRE. Any item submitted to the RIA repository is assigned a PID based on the Handle System, which does not have to be changed when the system migrates to new hardware, or when changes are made to the system, as appropriate to maintain the identifier integrity. Besides depositing scientific publications as an institutional minimum request, RIA will be used to deposit EPIBOOST policy briefs, academic theses and conference communications.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

● **Biblio – Institutional repository of the UGent** (https://biblio.ugent.be/)
All published research ouputs in which UGent is a co-author will be deposited in the institutional repository biblio. Depending on the policy of the journal, both the publisher's version as well as the author's accepted version will be uploaded in the repository. The author's accepted version can be legally made open access if the journal does not provide open access or open access fees are out of the foreseen budget. Biblio functions as a digital archive to preserve the University's scholarly publications and make them as widely available as possible through its open access infrastructure.

● **DIGITAL.CSIC – Institutional repository of CSIC** (https://digital.csic.es/).
Following the CSIC's open access mandate (http://digital.csic.es/dc/mandato-oa-csic.jsp), all generated publications will be available in the home CSIC institutional repository, DIGITAL.CSIC. We will ensure open access to all peer reviewed scientific publications relating to the results of the project by depositing a copy of the published version or final peer-reviewed manuscript accepted for publication therein. DIGITAL.CSIC is one main data provider at OpenAIRE, the European Commission aggregator of open science. Deposit onto DIGITAL.CSIC will be realized upon publication if an electronic version is available for free via the publisher, or within six months of publication in any other case. DIGITAL.CSIC ensures open access to the bibliographic metadata that identify the deposited publication.

4.2.2 Data

All research data that constitute EPIBOOST results will be made openly available through deposit in public/institutional repositories as described above. This includes DNA methylation datasets concerning different species and challenges experimentally imposed to the organisms; transcriptomic (gene expression) datasets concerning different species and challenges experimentally imposed to the organisms; phenotypic responses of challenged organisms, concerning several species and selected environmental contaminants (Appendix A).

Considering that these data are the basis for scientific publications (where the necessary statements pointing to the data availability will be added), and that both GEO and Zenodo allow it, submitted data will be embargoed until publications have been accepted, yet reviewers access will be allowed during the applicable editorial stages. Accessibility to the data by reviewers in GEO will be provided through a reviewer token allowing anonymous, read-only access to the corresponding private submissions after approval in the repository. This token is shared with editors handling each manuscript within the editorial pipeline and they will manage the access to the data according to the journal's privacy guidelines. In Zenodo, a request for access to data identified by a DOI in a manuscript submission can be made by editors and reviewers; the access upon request is handled by the data owner, thus access is controlled by EPIBOOST data owners always, through an automatically generated secret link. The time frame for this embargo cannot be accurately estimated at this stage since it depends on the time frame of editorial flows at each journal where EPIBOOST research is submitted. From publication onwards, data will be openly available, both directly via repositories and as integrated in scientific papers published under CC BY licences allowing immediate access via repository, in both cases retaining sufficient intellectual property rights of authors.

Data included in academic theses are a particular case concerning open sharing due to the educational nature. In this way, it is common that academic theses contain data that will be latter published within scientific articles. In such a case, EPIBOOST theses will be deposited in the RIA repository with a reasonable embargo period (generally one year) that allows the necessary time for the final results sourced by the data to become published is scientific journals of the field. Another particular case in research data regards nucleic acids quantity and quality records concerning different species, before and after exposure experiments (see Appendix A). These data are essential operational, allowing the

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

optimization of protocols and the definition of data analysis strategies. As such, these data will not constitute items entitled to open sharing per se, although they may be included in scientific publications as needed. Internal accessibility of the EPIBOOST researchers to these data will be ensured as detailed at the beginning of section 4.2.

Most of the data that will be collected/generated during the implementation of EPIBOOST capacitation activities will be used for internal processing, internal use in some cases (to implement improvements through the timeline) and project reporting purposes only: data concerning participants and evaluation of Advanced Courses; data concerning participants and evaluation of Summer Schools; data concerning participants and evaluation of Short Courses; data concerning participants and evaluation of the workshop on ethically compliant use of animals in research (see section 3.2.2 for details on these data and Appendix A for a summary). These data contain sensitive information, namely personal data and personal opinions that, although anonymised, can be considered sensitive as per the size of the groups that will be surveyed. As such, the access to this data will be restricted by the applicable ethical considerations as described in section 8 and will not be made openly available. The access to these data is controlled by restricting access to shared archiving folders to the team members having defined responsibilities in the collection and processing of each dataset, as well as by keeping the corresponding project deliverables (Deliverables 16, 20 and 21).

Still concerning the implementation of capacitation activities, the data collected/generated on views and opinions of participants in the Focus Group on the management of Twinning actions (see section 3.2.2 for details on these data and Appendix A for a summary) are a particular case. Restrictions as detailed above for similar data will apply during collection and processing stages. However, as the Focus Group initiates meetings and collaborative writing of the intended good practices guide on the management of Twining projects, restrictions will need to be alleviated. Access to anonymised data collected in surveys will be given to the participants of the focus group for processing and analysis, so that the data can be used in the elaboration of the guide. Furthermore, processed data will be ultimately incorporated in the guide produced, which is a public deliverable of EPIBOOST (Deliverable 22) that will be made openly available in Zenodo.

4.2.3 Metadata

The deposit of data as described above will necessarily render the associated metadata openly available as required by the Grant Agreement, facilitating mining and machine-based accessibility to the data. Moreover, the metadata will be available associated with the data for how long the data are kept in the long-term, and will migrate associated to the data, if for any reason such an operation becomes necessary to preserve the data open availability. There are currently no perspectives of a situation where data made available are removed from such a status, and if that happens, the EPIBOOST team will ensure a re-deposit of (meta)data in alternative repositories at least during the next 5 years after the end of the project; therefore, we see no need to confirm *a priori* the availability of metadata in such a scenario.

In the elected repositories, metadata are integrated under the data identifier (see section 4.1.1) and indexed in the repository's search engines immediately after publishing. Metadata harvesting by other servers (DataCite, PubMed, OpenAIRE) is also a standard procedure upon record validation and publishing in each repository, greatly promoting discoverability and accessibility. In another dimension, accessibility to the metadata will be facilitated by the use of non-proprietary file formats widely used in the disciplinary field, but still, the (open source) software needed to read the data will be summarised whenever necessary.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

4.3 Making data Interoperable

Inter-disciplinary interoperability of data will be achieved through the use of standardized and organized controlled vocabularies and ontologies will be used, the chosen repositories being compliant with this principle. The EPIBOOST project will not generate project specific ontologies or vocabularies, relying instead in well-established standards available for the disciplinary fields where the project develops.

In this context, we will use the BioPortal (https://bioportal.bioontology.org/) as a browser for searching the most appropriate ontologies applying to the generated research data within each specific field and corresponding metadata. The BioPortal is a FAIR compliant (FAIRsharing record at DOI 10.25504/FAIRsharing.4m97ah) library of biomedical ontologies and terminologies developed in Web Ontology Language (OWL), Resource Description Framework (RDF)(S), Open Biological and Biomedical Ontologies (OBO) format, Protégé frames and Rich Release Format. Community-specific ontologies harvested by the BioPortal will be used according to the features of each dataset. For example, obvious ontologies at this stage are Gene Ontology (GO – OBO format providing three Vocabularies for the annotation of gene products regarding their molecular function, cellular component, and biological role), Gene Expression Ontology (GEXO – OWL format for the domain of gene expression, integrating data from NCBI) and Epigenome Ontology (EGO – OWL format for integrative epigenome knowledge representation and data analysis). Less obvious at this stage are the ontologies that will be definitively adopted for phenotypic datasets and rich metadata that will be associated to genomic datasets, entailing information on species, biological systems, treatments, stressors and ecological context. We can at this stage identify the FAIR-compliant ENVO (Environment Ontology) standard (FAIRsharing record at DOI: 10.25504/FAIRsharing.azqskx) for use in EPIBOOST, in such a way that interoperability can be maximised within the scope of the project. In fact, ENVO is an expressive, community ontology supporting humans, machines, and semantic web applications in the understanding of environmental entities, which has been applied to genomic data (e.g., supporting the metadata checklists of the Genomics Standards Consortium), and currently consists of a fully-fledged ontology within the OBO Foundry and Library. Besides the BioPortal, a number of portals and ontology browsing interfaces harvest ENVO, including the European Bioinformatics Institute's Ontology Lookup Service (OLS). In what concerns chemicals, the standard IUPAC nomenclature or conventionally used abbreviations for compounds and/or groups of compounds.

At this point in the project timeline, establishing the cross-referencing of the data with other data from the project is difficult. However, by adhering to the FAIR principles, EPIBOOST will promote the inclusion of qualified references to other data as much as possible in the shared datasets, both embedded in the deposited documents and in associated metadata. The cross-referencing of datasets from previous research can be however established for our genomic datasets, which will definitively include reference genomes quotation in the associated metadata, retrievable via standard identifiers. Moreover, when raw data and processed data compose a deposited dataset (e.g., epigenomics and transcriptomics data), qualified references will be provided in the GEO submission documentation, promoting a Series record assigned an omnibus GEO assession number.

4.4 Increase data Re-use

Genomic and transcriptomic data will be deposited under compliance with minimal information standards such as MINSEQE, and relevant codes will be made available along with all relevant supportive documentation contextualizing and instructing for reuse using a standardized structure and language. Provisions detailed above for data and metadata findability and accessibility were established so that future free reuse of openly shared datasets is the widest possible. Phenotypic datasets will be deposited in Zenodo, hence complying with the best standards for facilitated widely

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

reuse; the same applies to the data underlying and included in the good practices guide for Twinning management.

Reusability of data will be automatically ensured by the overall deposit and licensing strategy. In general, research data sharing will rely in the CC0 Public Domain Dedication, added requests for credit (authorship, citation and/or acknowledgement details), to prevent ambiguity-derived reuse limitations. This applies to data deposited in the public repositories GEO and Zenodo. The EPIBOOST publications that use generated data will cross reference the datasets, but will be available openly under the latest available CC-BY licence, hence providing minimum intellectual property rights to the authors.

As per the EPIBOOST strategy of depositing data openly for long-term storage in Open Science compliant repositories, the corresponding research and capacitation (Guide) results will be useable by third parties after the end of the project, for the repositories lifetime; the elected public and institutional repositories have no foreseeable termination at least in the current moment in time. Concerning data that will not be openly shared externally, storage will be carried out in the EPIBOOST Onedrive platform (UAVR; see section 4.2) for at least five years of the project afterlife, and access will be allowed upon reasonable request by the EPIBOOST team members and beneficiary institutions.

Before depositing and publication, during internal data handling and processing stages, the responsibility for ensuring data quality is entitled primarily to the partner generating them (see section 3.2.4 and Appendix A for a summary); data provenance is unambiguously clarified by the file naming convention adopted in the project (see section 4.1.1). In this context, the PIs defined for each group of organisms addressed in EPIBOOST and the sequencing work within the UAVR team (WP1 leader) are the primary responsible for ensuring the quality and integrity of the data produced. However, due to the EPIBOOST nature, both CSIC and UGent play a supervision role in this context defined on the basis of task leadership in WP2 (hands-on training). Internal review of all datasets produced will be made at least by these aforementioned key players, besides ordinary cross-review procedures among team researchers directly collecting and handling the data.

5. Other research outputs

All developed bioinformatics workflows and scripts will be based on open-source software and packages (e.g., R and R packages commonly used for the analysis of DNA methylation and gene expression datasets). Given the fast evolution of bioinformatics software and development of novel algorithms, it is impossible to describe *a priori* all software and packages that will be used. Therefore, software and packages used will be recorded and a more feasible the list will be updated throughout the project as more packages and software will be integrated. All workflows and scripts will also be made available via deposit in GitHub (DOI: 10.17616/R3559G), an open-source developer's platform for software and bioinformatics that integrates with Zenodo for gaining the benefits associated with this repository (see section 4.2.1).

There are materials that will be reused or produced during EPIBOOST. These include: experimental protocols and sequencing protocols that have been developed previously by the partners, and will be adapted for tackling the changes inherent to the research with different species as implemented in WP1 species; training materials that reflect the expertise of the partners and compose courses and workshops that will be implemented during the project. These materials will be reused or generated as part of the capacitation ambitions of EPIBOOST and will therefore remain shared within the Consortium only. The storage and corresponding security aspects for these materials are as described

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

before and below for data that will not be deposited for open access in the long-term in public or institutional repositories.

6. Allocation of resources

Costs related to local data management, including storage, security and local archiving are covered by the budget of the project. Institutional services and support for implementation of local data management strategies are part of the indirect costs claimed by beneficiaries. In this context, each beneficiary generating or reusing research data will be responsible for their quality, organization and management, ensuring integrity and recoverability as necessary. According to the present Data Management Plan, these processes will require dedicated personnel effort and the associated claimed costs. The costs of data generation, processing, including validation and revision before deposit for long term preservation are considered per task as applicable, both as purchase and personnel costs (individual researchers have personal responsibility for the curation, validation and deposition of data generated under their coordination/supervision). Physical storage devices may be needed while processing the data, depending on the software used and the related needs. In such case, the acquisition of dedicated physical drives will be covered by purchase costs considered in the budget.

Other data management costs, especially those related to long-term preservation are not foreseen, considering that the elected repositories do not apply fees, while all ensure very large time frames of depositing operation (e.g., at least 20 years as claimed by Zenodo) or free-of-charge migration to suitable equivalent repositories if necessary. Considering that some data can be included in scientific publications, it is worth addressing any potential costs related to publication in open access (gold) under CC BY licences. Many publishers adhere to the EU plan S and have been enforcing agreements with institutions included in the b-On consortium, which is the case of the EPIBOOST beneficiaries. However, some publishers and some journals still have non-waived article processing charges. The UAVR budget considered this possibility and will be used to cover such costs if necessary.

Decisions concerning data management in general, and any necessary allocation of resources will be entitled to the research PIs at each institution, namely Joana Pereira at UAVR, Jana Asselman at UGent, and Laia Navarro Martin at CSIC. Depending on data provenance (see Appendix A), the corresponding PI will decide what data will be kept and how (e.g., the repository where data will be submitted for long term storage) yet complying with the provisions of the present DMP.

7. Data security

The three partners (UAVR, UGent, and CSIC) have their own reliable policies for data management and data security before data is placed into public repositories. All data will be stored and transferred according to applicable national, EU and international data security regulations and guidelines. Sensitive data will be handled and stored in the short-term, securely and according to the best ethical practice as detailed in section 8.

Local, short-term storage of data will be primarily made in the EPIBOOST Onedrive platform hosted in the UAVR DataCenter of the Information and Communication Technology (ICT) Services. This platform has controlled access and is based on cloud storage, with regular backups in place. Access

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

to the research data will be restricted to the team researchers designated for so by the project scientific coordinators at each partner institution and access permissions will be given accordingly by the overall scientific coordinator owning the Onedrive storage privileges. Permissions for access to sensitive data will be severely restricted to the short-list of responsible team members for the collection and the processing of these data, and will be given on the basis of each collection and handling event or associated reporting responsibilities. Recoverability of data will be ensured by doubling (at least) local storage, thus, besides the primary cloud storage, data will be stored in dedicated external hard drives; research data may as well be stored temporarily in their working computers/laptops provided their internet connections are institutional or secured by institutional VPNs. Non sensitive data and research data will be kept locally stored for five years after the end of the project, allowing medium-term reuse upon reasonable request within the Consortium and recovery as needed.

All elected public and institutional repositories for open sharing and long-term preservation of data and research outputs (see section 4.2.1) are all trusted for long-term preservation and curation as detailed previously in the present document. Data security is hence ensured for therein deposited data.

8. Ethics

Ethical issues impacting data sharing relate to compliance with European personal data protection policies. These are applicable to data collected for reporting purposes on the implementation of Advanced Courses, Summer Schools, Short-Courses and workshops, as well as on the establishment and activity of the Focus Group on the management of Twinning projects (see Appendix A for a summary).

Partner institutions collecting and handling these data (see origin/provenance in Appendix A) will adhere to local guidelines under the oversight of an institutional Data Protection Officer (DPO) and nominated Data Controller/Responsible (per dataset collection). The primary role of the DPO is to ensure that EPIBOOST processes the personal data of any stakeholders, participants in EPIBOOST activities or any other individuals are following the applicable data protection rules locally and the EU General Data Protection Regulation (Regulation [EU] 2016/679) regarding privacy, confidentiality and consent. DPOs have expert knowledge on data protection issues and associated ethics, and will be given good understanding of the project activities where data protection applies and linked project needs by the team members responsible for their implementation or by the scientific coordinators at each institution. On this basis, DPOs will ensure that data controllers/responsibles and subjects are informed about and aware of their data protection rights, obligations and responsibilities; provide advice and recommendations to the partners about the interpretation or application of the data protection rules; monitor the elaboration of data protection dossiers for each activity as needed; ensure data protection compliance within the project; handle queries or complaints on participants in EPIBOOST activities; cooperate and support the response to requests by the DPO; monitor the project compliance with the applicable data protection rules.

Within the contextual framework above, data collection will be mostly based on surveys set to promote or assess the efficiency of EPIBOOST activities. These surveys will be implemented in trustable and secure web-based platforms, preferably institutionally supported. For example, surveys carried out by UAVR will be carried out using the official forms platform used at the University of Aveiro, hosted in the DataCenter of the Information and Communication Technology (ICT) Services. The platform has

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

access control (login and password) and the existence of backups is guaranteed. The data collected by this form will be saved on UA's ICT services servers, with access control and automated backups, and will be kept after the end of the respective project activity for the minimum possible period of time considering their purposes, being available only for the data processing team within each activity. Participants will not be asked to provide personal data apart from aspects critical to the survey purposes and reporting requests (e.g., nationality, gender) but as a standard, such information will be anonymized to the largest possible extent. The informed consent applying to the data sharing needs will be requested in each survey, its specific terms being appropriately clarified to subjects in each survey. The same principles will apply for the collection of personal data in the form of images, sound and/or video during EPIBOOST activities for reporting and dissemination/communication purposes (e.g., social media coverage), with the appropriate contingency measures applying as necessary to prevent constraints in participation in EPIBOOST activities by subjects that do not authorise data collection (e.g., establishment of exclusion zones where this documentation will not be made within the functional space of the room where activities take place).

The storage of personal, anonymised and pseudoanonymised data will be supported by each partner collecting and handling the data, complying Regulation [EU] 2016/679 according to principles of proportionality and necessity. As a standard, these data will be kept up to 6 months after the end of the project EPIBOOST, will not be transferred to third parties; in what is tangible to external knowledge of these data, they will be used for project reporting to the funding agency and authorised dissemination purposes only.

9. Other issues

Among the three partners of the EPIBOOST Consortium, all have institutional guidelines for data management. These are compatible and complementary to the provisions of the present Data Management Plan.

10. Disclaimer

EPIBOOST has received funding as a Coordination and Support Action from the European Union under Horizon Europe (Grant no. 101078991). Views and opinions expressed in the present Data Management Plan are however those of the authors and the EPIBOOST Consortium in general only and do not necessarily reflect those of the EU or the European Research Executive Agency. Neither the EU nor the granting authority can be held responsible for them.

Re-use of information contained in this document for commercial and/or non-commercial purposes is authorized and free of charge, as long as the following conditions are met: acknowledgement by the re-user of the source of the document; non-distortion commitment of the re-user regarding the original meanings assumed in the original document; non-liability of the EPIBOOST Consortium and/or partners for any consequence stemming from the re-use. Moreover, it is noteworthy that the present document will be updated at least once during the EPIBOOST project timeline; updated version(s) may contain revisions, corrections, editions and extensions of the contents.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

Questions, comments or queries on or related to the present document and its contents can be addressed to the EPIBOOST Coordinator via email: cesam-epiboost@ua.pt.

EPIBOOST (www.epiboost.web.ua.pt) is supported by the European Commission under the Horizon Europe – Coordination and Support Actions Framework Programme, HE CSA, Grant Agreement number: 101078991

WP	Task	Data short description	Re-use or generated	Type	Formats	Size	Origin/ provenance
1	1.2 1.3	Nucleic Acids quantity and quality data concerning different species, before and after exposure experiments	generated	• Numerical • Quantitative • Raw	.xls/.xlsx, .csv, .pdf	200 MB	• Experimental • UAVR, CSIC, UGent
1,2	1.3 1.5 2.4	Genomes and/or transcriptomes of the species targeted in the project, with the corresponding metadata	Re-use	• Textual • Quantitative • Raw and processed	.txt, .gff, fq. fast5	60 GB	• NCBI, ENA
1,2	1.3 1.5 2.4	DNA methylation datasets concerning different species and challenges experimentally imposed to the organisms	generated	• Textual and numerical • Quantitative • Raw and processed	.txt, .csv, .dta, Fq., fast5	1.5 TB	• Experimental • UAVR, CSIC, UGent
1,2	1.3 1.5 2.4	Transcriptomic datasets concerning different species and challenges experimentally imposed to the organisms	generated	• Textual and numerical • Quantitative • Raw and processed	.txt, .csv, .dta, .fq, fast5	500 GB	• Experimental • UAVR, CSIC, UGent
1	1.4 1.5	Phenotypic responses of challenged organisms (several species and 2-3 environmental contaminants)	generated	• Textual, numerical, graphical • Quantitative • Raw and processed	.tif, .jpg, .xls/.xlsx, .pdf .doc/.docx, .txt, .csv	5 GB	• Experimental • UAVR, CSIC, UGent
3	3.1	Data concerning participants and evaluation of Advanced courses	generated	• Textual, numerical, graphical • Quantitative and Qualitative • Raw and processed	.xls/.xlsx, .csv, .doc/.docx, .tiff, .jpg, .pdf	500 MB	• Surveys • UAVR
3	3.2	Data concerning participants and evaluation of Summer Schools.	generated	• Textual, numerical, graphical • Quantitative and Qualitative • Raw and processed	.xls/.xlsx, .csv, .doc/.docx, .tiff, .jpg, .pdf	500 MB	• Surveys • CSIC, UGent
3	3.3	Data concerning participants and evaluation of short-courses	generated	• Textual, numerical, graphical • Quantitative and Qualitative • Raw and processed	.xls/.xlsx, .csv, .doc/.docx, .tiff, .jpg, .pdf	500 MB	• Surveys • UGent
4	4.1	Data on participants and evaluation of the workshop on ethically compliant use of animals in research	generated	• Textual, graphical • Quantitative and Qualitative • Raw and processed	.xls/.xlsx, .csv, .doc/.docx, .tiff, .jpg, .pdf	500 MB	• Surveys • UAVR
5	5.2	Contacts of coordinators of sister projects	Re-use	• Textual • Qualitative • Raw	.xls/.xlsx, .csv	100 KB	• EU-REA
5	5.5	Personal data, views and opinions of participants in the Focus Group on the management of Twinning actions	generated	• Textual, numerical, graphical • Quantitative and Qualitative • Raw and processed	.xls/.xlsx, .csv, .doc/.docx, .tiff, .jpg, .pdf	5 GB	• Surveys, meetings • UAVR